Breast Cancer Analysis Using ANN

Breast cancer is the most common cancer among women worldwide, accounting for 25% of all cancer cases. In 2015, over 2.1 million people were diagnosed with breast cancer. This project focuses on using machine learning techniques, particularly Artificial Neural Networks (ANNs), to classify tumors as either malignant (cancerous) or benign (non-cancerous) based on the Breast Cancer Wisconsin (Diagnostic) Dataset. By building and optimizing an ANN model, the goal is to develop an accurate classification system for breast cancer diagnosis.

About Dataset

The **Breast Cancer Wisconsin (Diagnostic) Dataset** contains 30 features derived from digitized images of fine needle aspirates (FNA) of breast masses. These features describe the characteristics of cell nuclei, such as radius, texture, perimeter, area, and compactness. The dataset includes 357 benign and 212 malignant cases, with no missing values. It provides crucial data for classifying tumors as benign or malignant, with each feature computed as mean, standard error, and worst (largest) values. The dataset is accessible via UCI Machine Learning Repository and other sources.

Link: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data/data

Data Cleaning

In the data cleaning process, we first checked for missing values using isnull().sum() to identify any columns with null values. We then dropped columns with more than 30% missing data. In this case, the column Unnamed: 32 was dropped because it exceeded the threshold of 30% missing values. After that, we checked for duplicate rows using the duplicated().sum() method and found that there were no duplicate entries in the dataset. The shape of the dataset was updated after the removal of the column, ensuring the data is clean and ready for further analysis.

EDA

We used a heatmap in our exploratory data analysis (EDA) to look at the relationships between different dataset elements. Finding characteristics that are highly associated is essential to preventing multicollinearity because the dataset has a lot of features. To identify duplicate traits, we visualized the associations between them using a correlation heatmap. We used feature selection methods like ANOVA F-value and Mutual Information to further minimize the dimensionality and concentrate on the most crucial characteristics for the model. By assisting in the identification of characteristics that exhibit the strongest correlation with the target variable, these techniques enhance the effectiveness and performance of the model.

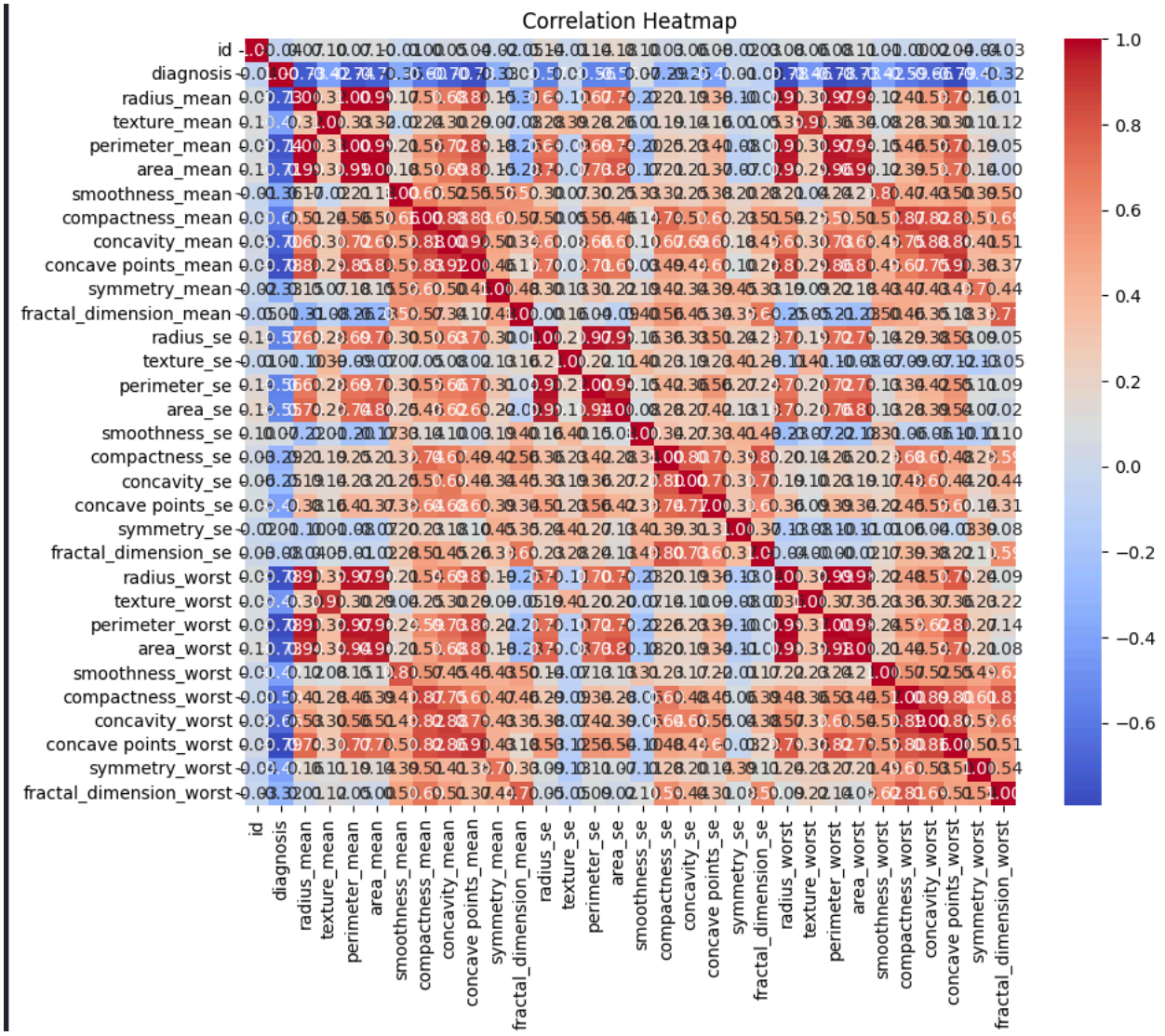

Fig: Correlation Matrix

During EDA, we observed that most numerical features were approximately normally distributed based on distribution plots. Since many machine learning models, including neural networks, perform better with standardized data, we decided to apply StandardScaler to scale the features, ensuring they have a mean of 0 and a standard deviation of 1.

We performed feature selection using ANOVA F-value and Mutual Information, identifying key features for predicting breast cancer diagnosis. We selected the top 10 features using ANOVA and trained an Artificial Neural Network (ANN) model with the MLPClassifier. After scaling the features with StandardScaler, the model was trained with two hidden layers (50 neurons each) and evaluated using classification metrics. The model achieved promising results, with a high

classification accuracy, precision, and recall, as indicated by the classification report and confusion matrix, demonstrating its effectiveness for predicting tumor malignancy.

Model 1: Base Model and using Feature Importance
The model achieved an outstanding accuracy of 97%, demonstrating its effectiveness in classifying breast cancer diagnoses. Precision and recall were high for both benign (B) and malignant (M) cases, with benign cases showing 99% recall and malignant cases achieving 95% recall. The confusion matrix reveals minimal misclassification, with only one benign case incorrectly classified as malignant and two malignant cases misclassified as benign. The F1-scores of 0.98 for benign and 0.96 for malignant further highlight the model's balanced performance. Training the model with both the top 10 features and the full dataset resulted in consistent, excellent performance, making the model highly reliable for cancer diagnosis predictions.

Model 2: Tuned Model

After performing hyperparameter tuning using GridSearchCV, the best model was identified with the following optimal parameters:

- **Activation function**: 'relu'
- **Solver**: 'adam'
- **Learning rate**: 'constant'
- **Alpha (regularization)**: 0.0001
- **Batch size**: 64
- **Hidden layer size**: (100,)

The best cross-validation score achieved was 97.8%. The performance of the tuned model on the test set was also excellent, with an accuracy of 97%. The classification report and confusion matrix show high precision and recall for both benign (B) and malignant (M) classes. The model misclassified only 1 benign and 2 malignant cases, reflecting its strong predictive power. This tuned model is robust and reliable for breast cancer diagnosis.

In conclusion, this project demonstrates the effective use of machine learning, specifically an Artificial Neural Network (ANN), to predict breast cancer diagnoses from the Breast Cancer Wisconsin (Diagnostic) dataset. After applying data preprocessing techniques and feature selection, we trained and evaluated the model using both the untuned and tuned versions. The results showed that there was not much difference in performance between the two models, with both achieving high accuracy (97%) and excellent precision and recall. Given this, we opted to use the untuned model for deployment, as it provides similar results while reducing the computational cost associated with training the tuned model.

Final Output: